

# A Dichotomy in Machine Knowledge

Samuel A. Alexander\*

*Department of Mathematics, the Ohio State University*

August 5, 2011

## Abstract

We show that a machine, which knows basic logic and arithmetic and basic axioms of knowledge, and which is factive (knows nothing false), can either know that it is factive, or know its own Gödel number, but not both.

## 1 Introduction

This is not a paper about artificial intelligence or conscious machines. But for motivational purposes, temporarily imagine we did have such a machine, whatever exactly that entails. We could ask this machine to tell us everything it knows: to enumerate its knowledge. It would then begin telling us various things: “ $1 + 1 = 2$ ”, “there are infinitely many primes”, and so on. It would also tell us things about its knowledge: “I know that  $1 + 1 = 2$ ”. We can further restrict our request, asking the machine to list only facts which it knows in the language of (say) Peano Arithmetic extended by a connective  $K$  for knowledge (formalized in Section 2): “I know  $1 + 1 = 2$ ” becomes “ $K(1 + 1 = 2)$ ”. We can at least say one thing about this list of knowledge: it is recursively enumerable (at least if we let the machine enumerate without outside disturbance).

Conscious or intelligent computers are beyond the scope of this paper, but the above shows how we can study *machine knowledge* anyway. Namely, we study recursively enumerable sets of formulas in a language which includes a modal operator for knowledge. In this paper, the language will be that of  $PA$  augmented by a unary knowledge operator  $K$ . We abuse language and identify a machine with its set of knowledge. For example, when we say a machine is “factive”, we mean that its set of known formulas are all true (in a background model). If we say a machine “knows Peano arithmetic”, we mean that the axioms of  $PA$  are in the r.e. set.

*Trivial Examples.* There is the know-nothing machine: we ask it to enumerate its knowledge and it lists nothing. This machine is vacuously factive. We can say it satisfies the schema  $K(\phi) \rightarrow \phi$ . The hypothesis,  $K(\phi)$ , means that  $\phi$

---

\*Email: alexander@math.ohio-state.edu

is among the list of known formulas. The conclusion  $\phi$  means  $\phi$  is really true. Again, there is the all-knowing machine: it lists every formula. This machine is not factive. It does not satisfy  $K(0 \neq 0) \rightarrow (0 \neq 0)$  (the hypothesis is true, the conclusion false). Again, there is the machine which knows exactly the consequences of  $PA$ . This machine is factive (if the background universe is  $\mathbb{N}$ ) but it does not know itself to be factive (the schema  $K(K(\phi) \rightarrow \phi)$  fails).

We are primarily concerned with machines which have the following properties. (which we call the *axioms of a knowing machine*).

- Knowledge of Tautology:  $K(\phi)$  whenever  $\phi$  is tautological.
- Knowledge Modus Ponens:  $K(\phi \rightarrow \psi) \rightarrow K(\phi) \rightarrow K(\psi)$ .
- Knowledge of Arithmetic:  $K(\phi)$  whenever  $\phi$  is an axiom of  $PA$ .
- Closure:  $K(\phi)$  whenever  $\phi$  is one of the above assumptions.
- Factivity:  $K(\phi) \rightarrow \phi$ . Everything known is true.

Each of these assumptions is plausible, requiring little of the machine in question. All are standard in epistemology. We are also interested in two additional properties, and the goal is to show that, together with the above basics, these properties are individually possible (for some  $e \in \mathbb{N}$ ) but are mutually inconsistent:

- Knowledge of Factivity:  $K(K(\phi) \rightarrow \phi)$ .
- Knowledge of having Gödel number  $e$ :  $K(K(\phi) \leftrightarrow \ulcorner \phi \urcorner \in W_e)$ , where “ $\ulcorner \phi \urcorner \in W_e$ ” abbreviates a canonical sentence expressing that the Gödel number of  $\phi$  in the  $e$ th r.e. set.

In [4], it was shown that a knowing machine cannot know its own Gödel number. However, this “implicitly” assumed Knowledge of Factivity. I say this requirement was “implicit” because it was not explicitly spelled out, but rather part of a closure requirement: in essence, in the above list of assumptions, “Closure” and “Factivity” were permuted.

## 2 A Very Simple Modal Logic

In this section we formalize quantified modal logic. There are many ways to do this. We take a very simple and weak approach. This approach is so weak that (unlike many treatments of modal logic) it does not actually depart from standard first-order logic. By taking such a weak approach to modal logic, we eliminate a lot of technical difficulty.

**Definition 1.** Suppose  $\mathcal{L}$  is a first-order language and  $K$  is a symbol not in  $\mathcal{L}$  (we will call  $K$  an *unary modal operator*). The first-order language  $\mathcal{L}(K)$  obtained by weakly extending  $\mathcal{L}$  by  $K$  is defined inductively as follows:

1. All the function, constant, and predicate symbols of  $\mathcal{L}$  are also in  $\mathcal{L}(K)$ .
2. For every formula  $\phi$  of  $\mathcal{L}(K)$ ,  $\mathcal{L}(K)$  contains a new 0-ary predicate symbol  $K_\phi$ .

**Notation 2.** Write  $K(\phi)$  for  $K_\phi$ . The abbreviation  $K(\phi)$  may be pronounced “ $\phi$  is known” or “I know  $\phi$ ” or (for brevity) “know  $\phi$ ”.

The notation is applied inductively. For example, we may write  $K(K(\phi))$  to abbreviate  $K_{K_\phi}$ .

*Warning.* Notation 2 does not play nicely with variable substitution. For instance  $K(x = y)(x|z)$  (the result of substituting  $z$  for  $x$ ) is  $K(x = y)$ , not  $K(z = y)$ . Failure to heed this warning can lead to philosophical paradoxes. In our treatment, the schemas  $(\forall x\phi) \rightarrow \phi(x|t)$  and  $(t_1 = t_2) \rightarrow \phi(x|t_1) \rightarrow \phi(x|t_2)$  are valid, and the Substitution Theorem holds: after all, we have not left classical first-order logic! These things can fail in some treatments of modal logic (see Shapiro [5]).

Hereafter, let  $\mathcal{L}(K)$  be the language of  $PA$  weakly extended by  $K$ .

**Definition 3.** By a *knowing entity* we mean an  $\mathcal{L}(K)$ -structure  $\mathcal{M}$  with universe  $\mathbb{N}$ , interpreting symbols of  $PA$  as usual, such that  $\mathcal{M}$  satisfies the axioms of a knowing machine from the Introduction. By a *knowing machine* we mean a knowing entity  $\mathcal{M}$  with the property that  $\{\ulcorner \phi \urcorner : \mathcal{M} \models K(\phi)\}$  is r.e.

Note that an entity (hence a machine) is completely determined by the formulas which it knows. In fact, it would be possible to reformulate Definition 3 and define entities and machines to be sets of formulas; this is what Carlson does [1] (pp. 59, 61). For our purposes, Definition 3 is more convenient.

### 3 An Inconsistency Result

William Reinhardt [4] showed that a knowing machine cannot simultaneously know that it is factive, and also know its own Gödel number (although everything was formalized in different ways). In this section we offer a streamlined proof of this result.

**Proposition 4.** Let  $e \in \mathbb{N}$  and let  $\Sigma$  be the set of axioms of a knowing machine together with Knowledge of Factivity and Knowledge of having Gödel number  $e$ . Then  $\Sigma$  is inconsistent.

*Proof.* By Gödel’s diagonal lemma, there is a sentence  $\phi$  such that Peano Arithmetic proves  $\phi \leftrightarrow \ulcorner \phi \urcorner \notin W_e$ .

Work in  $\Sigma_0 = PA \cup \{K(\phi) \rightarrow \phi, K(\phi) \leftrightarrow \ulcorner \phi \urcorner \in W_e\}$  (a subset of the consequences of  $\Sigma$ ). By PA, we have  $\phi \leftrightarrow \ulcorner \phi \urcorner \notin W_e$ . Combining this with  $K(\phi) \leftrightarrow \ulcorner \phi \urcorner \in W_e$ , we have  $K(\phi) \leftrightarrow \neg\phi$ . Assuming  $\neg\phi$ , we obtain  $K(\phi)$ , and then by  $K(\phi) \rightarrow \phi$ , we obtain  $\phi$ . Altogether, this establishes  $\phi$ .

I proved  $\phi$  from  $\Sigma_0$ . Thus there are finitely many axioms  $\sigma_1, \dots, \sigma_n$  from  $\Sigma_0$  such that  $\sigma_1 \rightarrow \dots \rightarrow \sigma_n \rightarrow \phi$  is a tautology. Now

$$\begin{aligned}
\Sigma &\models K(\sigma_1 \rightarrow \dots \rightarrow \sigma_n \rightarrow \phi) && \text{(Knowledge of Tautology)} \\
\Sigma &\models K(\sigma_1) \rightarrow \dots \rightarrow K(\sigma_n) \rightarrow K(\phi) && \text{(Knowledge Modus Ponens)} \\
\Sigma &\models K(\sigma_1) \wedge \dots \wedge K(\sigma_n) && (*) \\
\Sigma &\models K(\phi) && \text{(Modus Ponens)} \\
\Sigma &\models \ulcorner \phi \urcorner \in W_e. && \text{(Knowledge of having code } e)
\end{aligned}$$

Line  $(*)$  is true because for every element  $\sigma_i$  of  $\Sigma_0$ ,  $K(\sigma_i)$  is an axiom in  $\Sigma$ ; this is the only place where we invoke Knowledge of Factivity (one of the  $\sigma_i$  being an instance of Factivity). So  $\Sigma \models \phi$ ,  $\Sigma \models \ulcorner \phi \urcorner \in W_e$ , and  $\Sigma \models \phi \leftrightarrow \ulcorner \phi \urcorner \notin W_e$ , establishing inconsistency.  $\square$

**Theorem 5.** (*Reinhardt*) There is no knowing machine satisfying Knowledge of Factivity and Knowledge of having Gödel number  $e$ , for any  $e \in \mathbb{N}$ .

*Proof.* Such a machine would satisfy the set  $\Sigma$  from Proposition 4.  $\square$

## 4 Consistency of a machine knowing itself to be factive

In this section, we exhibit a knowing machine which possesses Knowledge of Factivity. We attempt to streamline Appendix B in Timothy Carlson's paper [1]. Carlson himself streamlined an argument due to Shapiro [5]. In both, Kleene's [3] *slash* operator was used; we short-circuit it.

**Definition 6.** Let  $\Sigma$  be the set of axioms of a knowing machine together with the axioms of Peano Arithmetic. Let  $\text{Slash}(\Sigma)$  be the  $\mathcal{L}(K)$ -model which has universe  $\mathbb{N}$ , interprets symbols of  $PA$  in the usual way, and interprets knowledge inductively as follows:

$$\text{Slash}(\Sigma) \models K(\phi) \text{ iff } \text{Slash}(\Sigma) \models \phi \text{ and } \Sigma \models \phi.$$

**Lemma 7.**  $\text{Slash}(\Sigma) \models \Sigma$ . Also,  $\text{Slash}(\Sigma)$  satisfies the schema  $K(K(\phi) \rightarrow \phi)$ .

*Proof.*

- (*Knowledge of Tautology*) If  $\phi$  is a tautology, then  $\text{Slash}(\Sigma) \models \phi$  and  $\Sigma \models \phi$ , so  $\text{Slash}(\Sigma) \models K(\phi)$ .
- (*Knowledge Modus Ponens*) Suppose that  $\text{Slash}(\Sigma) \models K(\phi \rightarrow \psi)$  and  $\text{Slash}(\Sigma) \models K(\phi)$ . This means  $\text{Slash}(\Sigma) \models \phi \rightarrow \psi$ ,  $\Sigma \models \phi \rightarrow \psi$ ,  $\text{Slash}(\Sigma) \models \phi$ , and  $\Sigma \models \phi$ . By Modus Ponens,  $\text{Slash}(\Sigma) \models \psi$  and  $\Sigma \models \psi$ . So  $\text{Slash}(\Sigma) \models K(\psi)$ .

- (*Knowledge of Arithmetic*) If  $\phi$  is an axiom of  $PA$ , then  $\text{Slash}(\Sigma) \models \phi$  since  $\text{Slash}(\Sigma)$  has universe  $\mathbb{N}$  and interprets symbols of  $PA$  in the usual way. Also,  $\Sigma \models \phi$  since  $\Sigma$  contains  $K(\phi)$  (Knowledge of Arithmetic) as well as  $K(\phi) \rightarrow \phi$  (Factivity). Altogether,  $\text{Slash}(\Sigma) \models K(\phi)$ .
- (*Closure*) If  $\phi$  is an instance of Knowledge of Tautology, Logic, or Arithmetic, then  $\text{Slash}(\Sigma) \models \phi$  by the above items. And certainly  $\Sigma \models \phi$ . So  $\text{Slash}(\Sigma) \models K(\phi)$ .
- (*Factivity*) If  $\text{Slash}(\Sigma) \models K(\phi)$ , then by definition  $\text{Slash}(\Sigma) \models \phi$ .
- (*Knowledge of Factivity*) Since  $\Sigma$  contains Factivity as an axiom,  $\Sigma \models K(\phi) \rightarrow \phi$ . We showed  $\text{Slash}(\Sigma) \models K(\phi) \rightarrow \phi$ . Together these show  $\text{Slash}(\Sigma) \models K(K(\phi) \rightarrow \phi)$ .

□

**Corollary 8.** For any  $\phi$ ,  $\text{Slash}(\Sigma) \models K(\phi)$  iff  $\Sigma \models \phi$ .

*Proof.*  $(\Rightarrow)$  By definition.  $(\Leftarrow)$  Suppose  $\Sigma \models \phi$ . By Lemma 7,  $\text{Slash}(\Sigma) \models \Sigma$ . Thus  $\text{Slash}(\Sigma) \models \phi$ . Together, this shows  $\text{Slash}(\Sigma) \models K(\phi)$ . □

**Theorem 9.** There is a knowing machine which possesses Knowledge of Factivity.

*Proof.* One such knowing machine is  $\text{Slash}(\Sigma)$ . It is a knowing entity by Lemma 7. It is a machine because  $\{\ulcorner \phi \urcorner : \text{Slash}(\Sigma) \models K(\phi)\} = \{\ulcorner \phi \urcorner : \Sigma \models \phi\}$  is r.e. □

## 5 Consistency of knowing one's own Gödel number

In this section we will construct a knowing machine which knows its own Gödel number. By Section 3, we cannot hope for such a machine to also know itself to be factive.

**Definition 10.** For every  $e \in \mathbb{N}$ , let  $\Sigma_e$  be the set of axioms of a knowing machine, along with Knowledge of having Gödel number  $e$ , the schema  $K(K(\phi) \leftrightarrow \ulcorner \phi \urcorner \in W_e)$ .

**Theorem 11.** There is an  $e \in \mathbb{N}$  such that there is a knowing machine which satisfies  $\Sigma_e$ . In words: there is a knowing machine with Knowledge of having Gödel number  $e$ .

*Proof.* Let  $\Sigma'_e$  consist of the axioms of  $PA$ , along with all the axioms of  $\Sigma_e$  except for Factivity, along with the schema  $K(\phi) \leftrightarrow \ulcorner \phi \urcorner \in W_e$  ( $\phi$  ranging over sentences).

For  $e \in \mathbb{N}$ , by a *coded consequence* of  $\Sigma'_e$ , I mean the Gödel number  $\ulcorner \phi \urcorner$  of a formula  $\phi$  such that  $\Sigma'_e \models \phi$ . Given  $e \in \mathbb{N}$ , we can effectively write a program to enumerate the coded consequences of  $\Sigma'_e$ . By the Church-Turing Thesis, there

is a total computable function  $f : \mathbb{N} \rightarrow \mathbb{N}$  such that for every  $e \in \mathbb{N}$ ,  $W_{f(e)}$  is the set of coded consequences of  $\Sigma'_e$ . By Kleene's Recursion Theorem, there is an  $e \in \mathbb{N}$  such that  $W_{f(e)} = W_e$ . Thus,  $W_e$  is the set of coded consequences of  $\Sigma'_e$ . Fix this  $e$  hereafter.

Let  $\mathcal{M}$  be the following  $\mathcal{L}(K)$ -structure. The universe of  $\mathcal{M}$  is  $\mathbb{N}$ , and symbols of PA are interpreted as usual. Predicate symbols are interpreted by  $\mathcal{M} \models K(\phi)$  iff  $\Sigma'_e \models \phi$ . I will show  $\mathcal{M} \models \Sigma_e$ , proving the theorem.

- (*Knowledge of Tautology*) If  $\phi$  is a tautology, then  $\Sigma'_e \models \phi$ , so  $\mathcal{M} \models K(\phi)$ .
- (*Knowledge Modus Ponens*) If  $\mathcal{M} \models K(\phi \rightarrow \psi)$  and  $\mathcal{M} \models K(\phi)$  then  $\Sigma'_e \models \{\phi \rightarrow \psi, \phi\}$ , thus  $\Sigma'_e \models \psi$ , so  $\mathcal{M} \models K(\psi)$ .
- (*PA*) With universe  $\mathbb{N}$  and interpreting symbols of PA as usual,  $\mathcal{M} \models PA$ .
- (*Closure, Knowledge of PA, Knowledge of having Gödel number  $e$* ) If  $\phi$  is an axiom of PA, an instance of  $K(\psi) \leftrightarrow \ulcorner \psi \urcorner \in W_e$ , or if  $K(\phi)$  is an instance of Closure, then  $\phi \in \Sigma'_e$  by construction, so  $\Sigma'_e \models \phi$  and  $\mathcal{M} \models K(\phi)$ .
- (*Having Gödel number  $e$* )  $\mathcal{M} \models K(\phi)$  iff  $\Sigma'_e \models \phi$ , iff  $\ulcorner \phi \urcorner$  is a coded consequence of  $\Sigma'_e$ , iff  $\ulcorner \phi \urcorner \in W_{f(e)} = W_e$ , which holds iff  $\mathcal{M} \models \ulcorner \phi \urcorner \in W_e$  (since  $\mathcal{M}$  has universe  $\mathbb{N}$  and interprets symbols of PA as usual).
- (*Factivity*) By the previous claims,  $\mathcal{M} \models \Sigma'_e$ . Assume  $\mathcal{M} \models K(\phi)$ . By definition,  $\Sigma'_e \models \phi$ . Since  $\mathcal{M} \models \Sigma'_e$ , we have  $\mathcal{M} \models \phi$ .

□

## References

- [1] Timothy J. Carlson (2000). Knowledge, machines, and the consistency of Reinhardt's strong mechanistic thesis. *Annals of Pure and Applied Logic* **105** 51–82.
- [2] Harvey Friedman and Michael Sheard (1987). An Axiomatic Approach to Self-Referential Truth. *Annals of Pure and Applied Logic* **33** 1–21.
- [3] Stephen C. Kleene (1962). Disjunctions and Existence under Implication in Elementary Intuitionistic Formalisms. *Journal of Symbolic Logic* **27** 11–18.
- [4] William N. Reinhardt (1986). Epistemic theories and the interpretation of Gödel's incompleteness theorems. *Journal of Philosophical Logic* **15** 427–474.
- [5] Stewart Shapiro (1985). Epistemic and intuitionistic arithmetic, in: Stewart Shapiro (Ed.), *Intensional Mathematics* (North-Holland, Amsterdam), pp. 11–46.